Asymptotic properties of sample quantiles from a finite population

Arindam Chatterjee

Abstract   In this paper we consider the problem of estimating quantiles of a finite population of size N on the basis of a finite sample of size n selected without replacement. We prove the asymptotic normality of the sample quantile and show that the scaled variance of the sample quantile converges to the asymptotic variance under a slight moment condition. We also consider the performance of the bootstrap in this case, and find that the usual (Efron's) bootstrap method fails to be consistent, but a suitably modified version of the bootstrapped quantile converges to the same asymptotic distribution as the sample quantile. Consistency of the modified bootstrap variance estimate is also proved under the same moment conditions.

Keywords   Quantile estimation · Finite population · Asymptotic normality · Variance estimation · Bootstrap

1 Introduction

Estimation of quantiles of a finite population has been an important problem in survey sampling. In this paper, we conduct a systematic and rigorous study of some important asymptotic properties of the sample quantile when a simple random sample is selected without replacement from a finite population. Most of the early work in finite population quantile estimation has been related to the construction of confidence intervals for population quantiles. For example, Woodruff (1952) considered the approach of using confidence intervals constructed by inverting the confidence intervals for the distribution function. Sedransk and Meyer (1978) considered exact confidence intervals using sample ordered statistics in case of stratified sampling. McCarthy (1965) and Smith and Sedransk (1983) studied lower bounds for confidence coefficients for

A. Chatterjee (✉)
Department of Statistics, Texas A&M University,
3143 TAMU, College Station, TX 77843, USA
e-mail: cha@stat.tamu.edu

stratified samples. Rigorous asymptotic results were derived by Francisco and Fuller (1991), who considered the quantile estimation problem under a complex sampling design. They showed the asymptotic normality of the estimated distribution function, the sample quantiles and also showed the consistency of Woodruff's method for constructing confidence intervals. Shao (1994) considered the problem of quantile estimation using the framework of L-statistics under a stratified multistage design and proved similar results as in Francisco and Fuller (1991), but under relatively weaker conditions. In both these papers due to the complexity of the survey designs used, the conditions imposed to prove the asymptotic normality are quite strong. In this paper we prove the asymptotic normality of sample quantile under some mild conditions, assuming a superpopulation model, where all probability statements have a clear interpretation, in terms of the underlying probability measure for the superpopulation. We also show variance of the sample quantile converges almost surely under this probability measure to the asymptotic variance under a mild moment and smoothness condition on the super-population distribution function. Another widely used approach to this problem involves using any available auxiliary information. Using the available auxiliary information, improved estimates of the distribution function are constructed, which in turn are inverted to give plausibly better estimates of the quantiles. More information on these methods can be found in the works of Chambers and Dunstan (1986), Rao et al. (1990), Mak and Kuk (1993) and the references therein. A Bayesian approach to quantile estimation from a finite population is considered in Nelson and Meeden (2006).

Variance estimation is one of the most important issues in survey research. When the parameter of interest is a nonsmooth function like a population quantile, the usual variance estimation methods like linearization, jackknifing or balanced repeated replication (BRR) either fail or are difficult to implement. In this situation the performance of the bootstrap (Efron 1979) is of interest. In this paper we also consider the performance of the bootstrap method that has been suggested in the finite population context by Gross (1980). In both cases, the limiting sampling fraction f is allowed to take values in [0, 1). We show that the usual version of the bootstrap fails if $f > 0$, i.e., if the sample size grows as a nontrivial fraction of the population size. We also propose a modification and show that the modified version of the bootstrap works for all values of $f \in [0, 1)$. Specifically we show that the cdf of the modified bootstrap quantile converges to the same asymptotic normal distribution and the bootstrap estimator of the variance of the sample quantile also converges in probability to the asymptotic variance.

The rest of the paper is organized as follows. In Sect. 2, we establish asymptotic normality of the sample quantile. In Sect. 3, we consider properties of the bootstrap. In Sect. 4 we carry out a simulation study to assess the finite sample performance of the modified bootstrap method. Proofs of all results are given in Sect. 5.

2 Main results

In order to have a suitable theoretical framework for developing asymptotic results in case of sampling from a finite population, it is common to assume that the finite

population is a random sample from a superpopulation (see Isaki and Fuller 1982). We assume that the superpopulation has an absolutely continuous cdf F_0 with density f_0. Also denote the underlying probability measure of the superpopulation as $\mathbf{P}_0$ and its corresponding expectation and variance as $\mathbf{E}_0$ and $\mathbf{V}_0$. Let r be a positive integer valued variable, i.e. r takes values in $\mathbb{N} = \{1, 2, \ldots\}$. We suppose that the population size N and the sample size n, are indexed by r and grows to infinity. For each $r \in \mathbb{N}$, let the finite population $\mathcal{X}_{N_r} = \{X_1, \ldots, X_{N_r}\}$ be an *iid* sample of size N_r selected from the superpopulation F_0. We select a simple random sample *without replacement* of size n_r, $\mathcal{Y}_{n_r} = \{Y_1, Y_2, \ldots, Y_{n_r}\}$ from the finite population $\mathcal{X}_{N_r}$. The population and sample cdf's are defined as,

$$F_{N_r}(t) = \frac{1}{N_r} \sum_{j=1}^{N_r} \mathbf{1}(X_j \le t), \quad t \in \mathbb{R},$$

and

$$\widehat{F}_{n_r}(t) = \frac{1}{n_r} \sum_{j=1}^{n_r} \mathbf{1}(Y_j \le t), \quad t \in \mathbb{R},$$

respectively, where $\mathbf{1}(A)$ denotes the indicator function for a set A. The corresponding *p-th* sample quantiles are $\xi_{p,N_r} = F_{N_r}^{-1}(p)$ and $\widehat{\xi}_{p,n_r} = \widehat{F}_{n_r}^{-1}(p)$. For all $r \in \mathbb{N}$, define the standardized sample quantile as,

$$U_r = a_r^{-1} \left(\widehat{\xi}_{p,n_r} - \xi_{p,N_r} \right),$$

where, $f_r = \frac{n_r}{N_r}$ is the sampling fraction and $a_r = \sqrt{\frac{1-f_r}{n_r}}$ is the norming constant. The sampling distribution function of U_r conditional on the population $\mathcal{X}_{N_r}$, is defined as

$$G_{n_r}(t) = \mathbf{P}_{\cdot|\mathcal{X}_{N_r}} (U_r \le t), \quad t \in \mathbb{R},$$

where $\mathbf{P}_{\cdot|\mathcal{X}_{N_r}}$ denotes the conditional probability distribution given the population $\mathcal{X}_{N_r}$. Similarly $\mathbf{E}_{\cdot|\mathcal{X}_{N_r}}$ and $\mathbf{V}_{\cdot|\mathcal{X}_{N_r}}$ will denote the corresponding expectation and variance. The first result gives conditions for the asymptotic normality of U_r.

Theorem 1 *Let $\xi_p = F_0^{-1}(p)$ be the p-th superpopulation quantile $(0 < p < 1)$. Assume that $f_0(\xi_p) > 0$ and that*

$$\lim_{r \to \infty} f_r = f \quad \text{for some } f \in [0, 1). \tag{1}$$

Then, $U_r \xrightarrow{d} N(0, \rho^2)$, a.s. $\mathbf{P}_0$, i.e.

$$\sup_{t \in \mathbb{R}} \left| \mathbf{P}_{\cdot|\mathcal{X}_{N_r}} (U_r \le t) - \Phi(t/\rho) \right| \to 0, \quad \text{as } r \to \infty, \text{ a.s. } \mathbf{P}_0,$$

where $\rho^2 = \frac{p(1-p)}{f_0^2(\xi_p)}$.

The next result gives conditions under which $\mathbf{V}_{\cdot | \mathcal{X}_{N_r}}(U_r)$ converges to the asymptotic variance ρ^2.

Theorem 2 *Assume that, there exists an $\alpha > 0$ such that $\mathbf{E}_0|X|^\alpha$ is finite* (*where the expectation is with respect to the superpopulation distribution F_0*). *Assume that the condition* (1) *of Theorem* 1 *holds. Then for any $\delta \in (0, \infty)$*

$$\sup_{r \geq 1} \mathbf{E}_{\cdot | \mathcal{X}_{N_r}} \left(|U_r|^{2+\delta} \right) < \infty \quad a.s. \ \mathbf{P}_0. \tag{2}$$

This theorem gives condition under which the sequence $\{|U_r|^k\}_{r \geq 1}$ is uniformly integrable for any $k \in (0, \infty)$ and in particular for $k = 2$. Hence the above two theorems imply variance consistency, which we state as a corollary.

Corollary 1 *Assume that the conditions in Theorems* 1 *and* 2 *are satisfied. Then*

$$\mathbf{V}_{\cdot | \mathcal{X}_{N_r}}(U_r) \to \rho^2 \quad as \ r \to \infty, \quad a.s. \ \mathbf{P}_0.$$

3 Bootstrap

The commonly used methods of obtaining variance estimates for nonlinear statistics in finite population sampling are linearization (taylor series) method and resampling techniques like the jackknife method, balanced repeated replications (BRR) and the bootstrap. The linearization method is inapplicable in case of quantiles. The usual jackknife method is known to have problems in case of nonsmooth functionals like quantiles (see Shao and Wu 1989). BRR works well for quantiles (Shao and Wu 1992) but they need careful construction of balanced subsamples, which is difficult even for slightly complex designs. Another method suggested by Woodruff (1952) involves using the estimated (normal) confidence intervals for the population quantiles, in order to obtain variance estimates for the sample quantile. Sitter (1992) studied the performance of Woodruff's method under stratified sampling using simulation based results and its performance was not satisfactory in some situations (also see Kovar et al. 1988). Other methods for variance estimation involve using plugin estimates of the asymptotic variance by estimating the superpopulation density at the population quantile (Shao 1994). Among all these variance estimation methods, the bootstrap is the most intuitively appealing method, due to its flexibility and also because it is applicable to both smooth and nonsmooth statistics. In an important work, Gross (1980) proposed using the bootstrap for estimating the variance of the sample quantile in the finite population context. Although Gross (1980)'s work has been extensively referred to the in the subsequent literature, the performance of bootstrap in estimating the asymptotic variance of the quantile seems does not seem to be known (see Sitter 1992; Shao and Chen 1998). To overcome the inadequacy of the usual (Efron's) bootstrap procedure in finite population setting different modifications like bootstrap without replacement, the use of different bootstrap sample sizes, rescaling bootstrap (Rao and Wu 1988) were suggested (see Shao 2003, for more details). But, none of these methods provide a clear answer in the case of sample quantiles.

In this paper we show that usual version of the bootstrap fails whenever $f > 0$. However with proper rescaling, the bootstrapped quantile converges to the same asymptotic normal distribution as the sample quantile for all $f \in [0, 1)$. We also show that the bootstrap variance estimate converges in probability to the asymptotic variance also, for all $f \in [0, 1)$. Thus, one can use the modified bootstrap estimator of this paper to estimate the variance of the sample quantile in a finite population setting. Related results on the use of bootstrap methods for variance estimation in finite population setup can be found in Bickel and Freedman (1984), Rao and Wu (1988) and Sitter (1992). Booth et al. (1994) construct a pseudo-population in the lines of a method suggested by Gross (1980), and bootstrap this psuedo-population to construct second order accurate confidence intervals for smooth functions of sample means. Shao and Chen (1998) establish the consistency of bootstrap procedures for sample quantiles when the dataset contains nonrespondents, which are imputed using hot-deck imputation. Lahiri (2003) gives a good review about the impact and use of bootstrap methods in survey research.

For the sake of completeness, we give a brief description of the bootstrap. Let $\mathcal{Y}_{n_r}^* = \{Y_1^*, \ldots, Y_{n_r}^*\}$ denote a bootstrap sample of size n_r, selected with replacement from $\mathcal{Y}_{n_r}$. The empirical cdf for the bootstrap sample is

$$\mathbf{F}_{n_r}^*(t) = \frac{1}{n_r} \sum_{j=1}^{n_r} \mathbf{1}(Y_j^* \le t), \quad t \in \mathbb{R}.$$

The pth bootstrap quantile is defined as $\xi_{p,n_r}^* = \mathbf{F}_{n_r}^{*-1}(p)$. The bootstrap version of U_r conditioned on the sample $\mathcal{Y}_{n_r}$ is defined as

$$U_r^* = a_r^{-1} \left(\xi_{p,n_r}^* - \widehat{\xi}_{p,n_r} \right).$$

The next result shows that this version of the bootstrap fails for all $f \in (0, 1)$.

Theorem 3 *Suppose that condition* (1) *of the Theorem* 1 *holds. Then,*

$$U_r^* \xrightarrow{d} N \left(0, \frac{\rho^2}{(1-f)} \right) \quad \text{in probability } \mathbf{P}_0.$$

Intuitively, this naive bootstrap fails because it does not account for the dependence in the sample $\mathcal{Y}_{n_r}$, which is obtained by sampling without replacement, while $\mathcal{Y}_{n_r}^*$ is obtained by sampling with replacement. However resampling without replacement is not feasible, as there is only one without replacement sample of size n_r from $\mathcal{Y}_{n_r}$, *viz.* itself. A simple fix for this problem is to rescale and define the new bootstrap version of U_r as

$$U_r^{**} = b_r^{-1} \left(\xi_{p,n_r}^* - \widehat{\xi}_{p,n_r} \right),$$

where $b_r^{-1} = a_r^{-1}\sqrt{1-f}$. The reason for re-scaling the bootstrap version of U_r is to make the bootstrap version consistent in estimating the asymptotic variance. The corresponding bootstrap cdf is defined as,

$$G_{n_r}^*(t) = \mathbf{P}_* \left(U_r^{**} \le t \right), \quad t \in \mathbb{R},$$

where, $\mathbf{P}_*$ is the bootstrap probability distribution conditioned on $\mathcal{Y}_{n_r}$. Similarly $\mathbf{E}_*$ and $\mathbf{V}_*$ will denote the bootstrap expectation and variance. The next result is about the asymptotic normality of U_r^{**} and consistency of the bootstrap variance estimator in estimating the asymptotic variance.

Theorem 4 *Assume that condition* (1) *in Theorem* 1 *holds. Then*

(a) $U_r^{**} \xrightarrow{d} N(0, \rho^2)$ *in probability* $\mathbf{P}_0$, *where* $\rho^2 = \frac{p(1-p)}{f_0^2(\xi_p)}$.

(b) *If we assume that* $\mathbf{E}_0|X|^\alpha < \infty$ *for some* $\alpha > 0$, *then* $\mathbf{V}_*(U_r^{**}) \to \rho^2$ *in probability* $\mathbf{P}_0$.

4 Numerical results

In this section, we conduct a small simulation study to compare the performances of the usual and modified (rescaled) bootstrap methods. A finite population of size N is chosen from a superpopulation F_0. Our main interest is to estimate the asymptotic variance ρ^2 (cf. Theorem 1) of the population median ($p = 1/2$). The asymptotic results derived in the previous sections will need to be used in a slightly modified manner for purpose of this simulation. Since the limiting sampling fraction f is unknown in a practical situation, hence the norming constant b_r is effectively taken as $\sqrt{n_r}$. Note that the limiting value of $f \in [0, 1)$ (cf. 1) allows such a modification.

For the simulation study we choose $F_0 \equiv N(0, 1)$, two population sizes of $N = 300$ and 1200 are selected. The sampling fraction $f \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For each choice of N and f, the initial finite population $\mathcal{X}_N$ (selected at random from F_0) is kept fixed during the simulation, which is expected in a finite population setup. From this finite population, M sets of samples of size $n = fN$ are selected without replacement, with the kth set denoted as $\mathcal{Y}_{n,k}; k = 1, \ldots, M$. Within each sample $\mathcal{Y}_{n,k}$, the sample p-quantile $\widehat{\xi}_{p,n}(k)$ is found. From each sample $\mathcal{Y}_{n,k}$, B bootstrap samples are selected, with the jth bootstrap sample denoted as $\mathcal{Y}_{n,k,j}^*; j = 1, \ldots, B$. The bootstrap quantiles in each bootstrap sample are denoted as $\xi_{p,n}^*(k, j)$. The corresponding version of U_r^* is

$$U_{k,j}^* = \sqrt{\frac{n}{1-f}} \left(\xi_{p,n}^*(k, j) - \widehat{\xi}_{p,n}(k) \right), \quad k = 1, \ldots, M, \ j = 1, \ldots, B,$$

and similarly $U_{k,j}^{**} = \sqrt{n} \left(\xi_{p,n}^*(k, j) - \widehat{\xi}_{p,n}(k) \right)$, are defined. The B bootstrap samples are used to construct a Monte-Carlo (MC) approximation of the bootstrap variance estimates. Corresponding to each sample $\mathcal{Y}_{n,k}$ the MC approximation of the usual bootstrap variance estimate is

Table 1 Simulation results comparing the estimated MSE and bias of the usual and modified (rescaled) bootstrap variance estimators at $N = 300$ and different values of f (or n)

f	$\widehat{\text{MSE}}\left(\widehat{\sigma^2}*\right)$	$\widehat{\text{MSE}}\left(\widehat{\sigma^2}**\right)$	$\widehat{\text{bias}}\left(\widehat{\sigma^2}*\right)$	$\widehat{\text{bias}}\left(\widehat{\sigma^2}**\right)$
0.1	2.0263	1.4554	0.7449	0.5133
0.3	1.8090	0.5156	0.8985	0.1577
0.5	6.8688	0.5373	2.2876	0.3584
0.7	19.1117	0.1506	4.2113	0.1638
0.9	222.1286	0.0819	14.6348	0.0497

Here $p = 0.5$, $M = 1800$ and $B = 500$. The true variance is $\rho^2 = 1.5707$

Table 2 Simulation results comparing the estimated MSE and bias of the usual and modified (rescaled) bootstrap variance estimators at $N = 1200$ and different values of f (or n)

f	$\widehat{\text{MSE}}\left(\widehat{\sigma^2}*\right)$	$\widehat{\text{MSE}}\left(\widehat{\sigma^2}**\right)$	$\widehat{\text{bias}}\left(\widehat{\sigma^2}*\right)$	$\widehat{\text{bias}}\left(\widehat{\sigma^2}**\right)$
0.1	0.7734	0.5566	0.3342	0.1437
0.3	0.7162	0.2167	0.5399	-0.0932
0.5	1.1326	0.2316	0.8509	-0.3599
0.7	12.2376	0.0666	3.4009	-0.0792
0.9	165.8798	0.0406	12.7916	-0.1345

Here $p = 0.5$, $M = 1800$ and $B = 800$. The true variance is $\rho^2 = 1.5707$

$$\widehat{\sigma^2}*(k) = \frac{1}{B} \sum_{j=1}^{B} \left(U_{k,j}^* - \frac{1}{B} \sum_{j=1}^{B} U_{k,j}^* \right)^2, \quad k = 1, \ldots, M.$$

Similarly $\widehat{\sigma^2}**(k)$ can be defined for the modified bootstrap version. In order to judge the accuracy of these estimates we make use of the M sets of samples $\mathcal{Y}_{n,k}$. This gives idea about the average performance of these estimates for a fixed finite population $\mathcal{X}_N$. Using these M values, a MC approximation of the MSE of these bootstrap variance estimates can be provided:

$$\widehat{\text{MSE}}\left(\widehat{\sigma^2}*\right) = \frac{1}{M} \sum_{k=1}^{M} \left(\widehat{\sigma^2}*(k) - \rho^2\right)^2,$$

and similarly for $\widehat{\sigma^2}**$. For comparison, we also include the MC estimate of the bias of the bootstrap variance estimates (denoted as $\widehat{\text{bias}}(\widehat{\sigma^2}*)$), which can be similarly defined as the MSE above. Tables 1 and 2 show the estimated MSE and bias of the usual and modified bootstrap variance estimates at population sizes $N = 300$ and 1200, respectively.

From the values in Tables 1 and 2 it is clear that the overall performance of the modified bootstrap variance estimate is far superior to the usual one. As the sampling

Usual bootstrap variance estimate   **Rescaled bootstrap variance estimate**
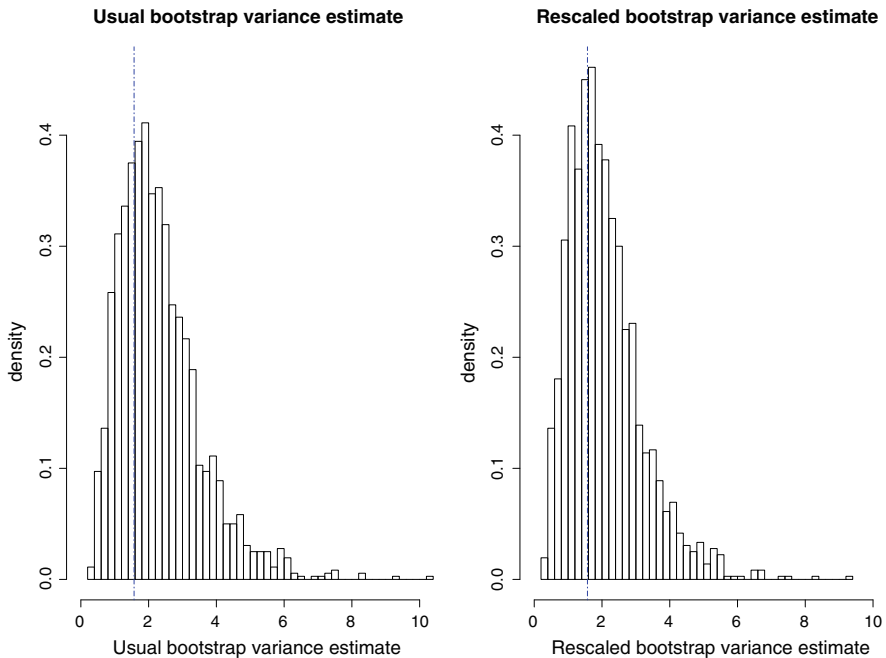
Fig. 1 $N = 300$, $f = 0.1$ and $p = 0.5$: Figure comparing the relative performance of the usual and modified (rescaled) bootstrap variance estimates. Here $M = 1800$ and $B = 500$. The *vertical dotted line* represents the true asymptotic variance $\rho^2 = 1.5707$

fraction f increases, the usual bootstrap performs worse with gross over estimation (also see Figs. 1, 2, 3). The extreme differences in the estimated MSE and bias of the usual and modified bootstrap variance estimates can be explained easily by observing that $\widehat{\sigma^{2*}}(k) = (1 - f)^{-1}\widehat{\sigma^{2**}}(k)$, for all k and for every choice of $f \in [0, 1)$. This implies that as f increases, the usual bootstrap variance estimate becomes more inflated due to the $(1 - f)$ term in the denominator. For example, in Table 1, at $f = 0.9$, $\widehat{\text{bias}(\sigma^{2*})} + \rho^2 = 16.205$ and $\widehat{\text{bias}(\sigma^{2**})} + \rho^2 = 1.6204$, which shows that the usual bootstrap estimate has approximately $(1 - 0.9)^{-1}(= 10)$ times more bias compared to the modified bootstrap estimate. The rescaling used in the modified bootstrap variance estimator overcomes this problem and instead of inflating the estimates, makes it more accurate as f (or n) increases. As N increases, the comparable performance at same f values is also better as expected. There is a tendency of slight under estimation by the modified bootstrap estimator at $N = 1200$, but it still performs better than the usual bootstrap estimate. It should be noted that the choice of the number of bootstrap samples B plays an important part in the simulation setup. The histograms in Figs. 1, 2, 3, compare the distributions of the two types of bootstrap variance estimates at $N = 300$ and different values of f. Though the modified bootstrap quantile performs well in this simple sampling scheme, it is desirable to study the performance of the bootstrap under more complex sampling designs.

Usual bootstrap variance estimate   **Rescaled bootstrap variance estimate**
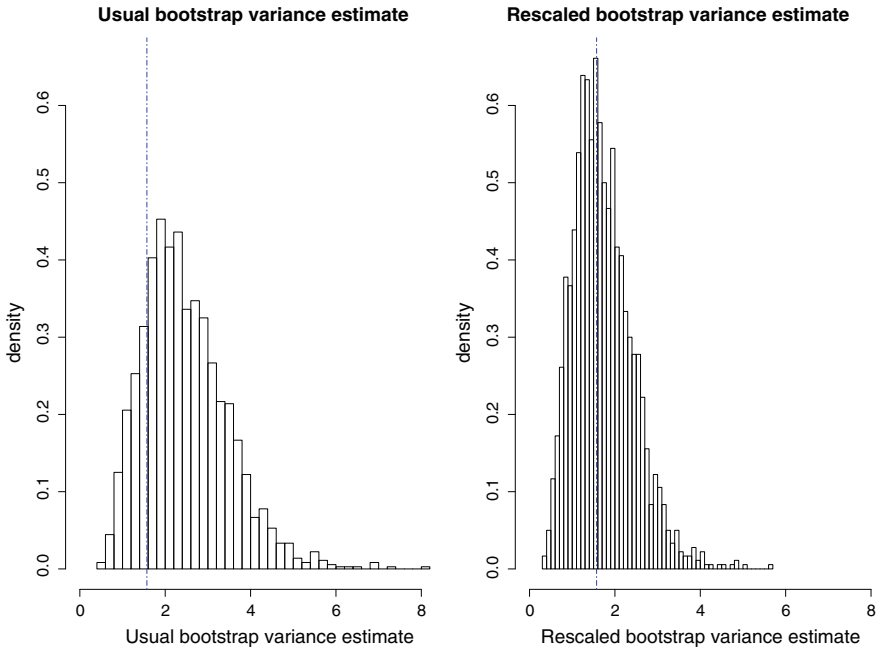
Fig. 2 $N = 300$, $f = 0.3$ and $p = 0.5$: Figure comparing the relative performance of the usual and modified (rescaled) bootstrap variance estimates. Here $M = 1800$ and $B = 500$. The *vertical dotted line* represents the true asymptotic variance $\rho^2 = 1.5707$

5 Proofs

5.1 Auxiliary results

We state the following Lemma's that will be used in our proofs.

Lemma 1 (Singh 1981, Lemma 3.1) *Let* $\{V_1, \ldots, V_k\}$ *are iid with* $V_i = 1 - a$ *or* $-a$ *with respective probabilities* a *and* $1 - a$, *then for any* $k \leq K$, $a \leq B$, $Z \leq D$ *with* $ZKB \leq D^2$ *we have*

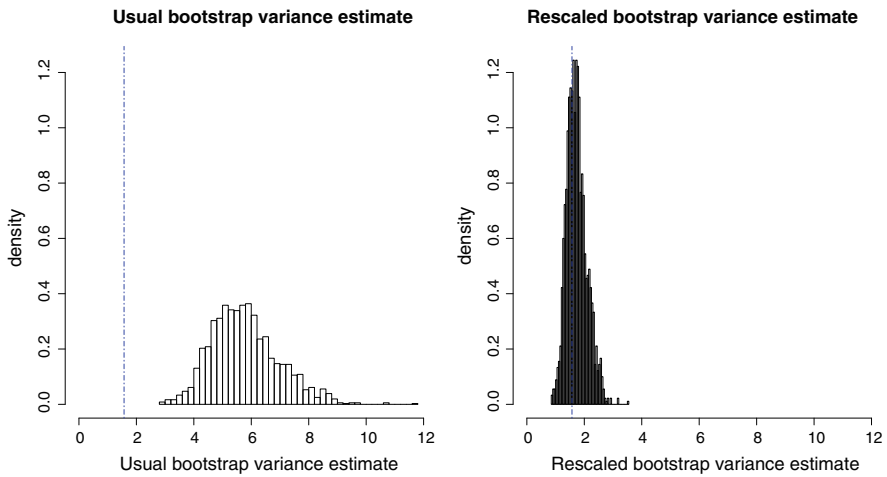$$\mathbf{P}\left(\left| \sum_{i=1}^{k} V_i \right| \geq (1 + e/2)D \right) \leq 2e^{-Z}. \tag{3}$$

Proof See Singh (1981).   □

Lemma 2 *For any fixed* $t \in \mathbb{R}$,

$$F_{N_r}(\xi_{p,N_r} + t a_r) - F_{N_r}(\xi_{p,N_r}) - F_0(\xi_{p,N_r} + t a_r) + F_0(\xi_{p,N_r}) = o(a_r), \tag{4}$$

almost surely ($\mathbf{P}_0$).

Usual bootstrap variance estimate   **Rescaled bootstrap variance estimate**

Fig. 3 $N = 300$, $f = 0.7$ and $p = 0.5$: Figure comparing the relative performance of the usual and modified (rescaled) bootstrap variance estimates. Here $M = 1800$ and $B = 500$. The *vertical dotted line* represents the true asymptotic variance $\rho^2 = 1.5707$

Proof Using the law of iterated logarithm for sample quantiles (Bahadur 1966) we can write,

$$\xi_{p,N_r} = \xi_p + O\left(N_r^{-1/2}(\log\log N_r)^{1/2}\right) \ a.s. \ (\mathbf{P}_0).$$

Define the quantity

$$\Delta_{N_r}(y) \equiv F_{N_r}\left(\xi_p + \frac{y}{\sqrt{N_r}} + ta_r\right) - F_{N_r}\left(\xi_p + \frac{y}{\sqrt{N_r}}\right) + F_0\left(\xi_p + \frac{y}{\sqrt{N_r}}\right)$$

$$-F_0\left(\xi_p + \frac{y}{\sqrt{N_r}} + ta_r\right).$$

Also define $\Delta^*_{1,N_r} = \sup\{\Delta_{N_r}(y) : |y| \leq \log N_r\}$ and $\Delta^*_{2,N_r} = \inf\{\Delta_{N_r}(y) : |y| \leq \log N_r\}$. In order to prove (4) it will be enough to show that

$$\Delta^*_{i,N_r} = o(a_r) \quad a.s. \ (\mathbf{P}_0) \quad \text{for } i = 1, 2.$$

Consider the intervals $B_i = (\frac{i-1}{\sqrt{N_r}}, \frac{i}{\sqrt{N_r}}]$, where $|i| = 0, 1, \ldots, (\lfloor\sqrt{N_r}\log N_r\rfloor + 1)$. So for any $y \in B_i$, we can write

$$\Delta_{N_r}(y) \leq \Delta_{N_r}\left(\frac{i}{\sqrt{N_r}}\right) + H_{N_r}(i) + f_0(\theta_1)\frac{1}{N_r} + f_0(\theta_2)\frac{1}{N_r},$$

where,

$$H_{N_r}(i) = F_{N_r}\left(\xi_p + \frac{i}{N_r}\right) - F_0\left(\xi_p + \frac{i}{N_r}\right) - F_{N_r}\left(\xi_p + \frac{i-1}{N_r}\right)$$
$$+ F_0\left(\xi_p + \frac{i-1}{N_r}\right),$$

with $\theta_1 \in (\xi_p + ta_r + \frac{i-1}{N_r}, \xi_p + ta_r + \frac{i}{N_r})$ and $\theta_2 \in (\xi_p + \frac{i-1}{N_r}, \xi_p + \frac{i}{N_r})$ (are obtained by using the mean-value theorem). Let,

$$M = \sup\{f_0(y) : |y - \xi_p| < \delta_0\}$$

be the supremum of f_0 in some δ_0-neighbourhood of ξ_p. We can similarly obtain a lower bound on $\Delta_{N_r}(y)$, so that for all $y \in B_i$ and for large enough r,

$$\Delta_{N_r}\left(\frac{i-1}{\sqrt{N_r}}\right) - H_{N_r}(i) - \frac{2M}{N_r} \le \Delta_{N_r}(y) \le \Delta_{N_r}\left(\frac{i}{\sqrt{N_r}}\right) + H_{N_r}(i) + \frac{2M}{N_r}.$$

Initially we only consider the upper bound for $\Delta_{N_r}(y)$ in the above inequality. Define, $R_{N_r} \equiv \max\{\Delta_{N_r}(\frac{i}{\sqrt{N_r}}) : |i| \le (\lfloor\sqrt{N_r}\log N_r\rfloor + 1)\}$. Then,

$$\Delta^*_{1,N_r} \le R_{N_r} + \max_{|i| \le (\lfloor\sqrt{N_r}\log N_r\rfloor + 1)} H_{N_r}(i) + \frac{2M}{N_r}.$$

Now choose a sequence $\{\epsilon_r : r \ge 1\}$, so that,

$$\frac{M}{N_r} \ll \epsilon_r a_r \ll M, \quad \text{and} \quad \epsilon_r \to 0. \tag{5}$$

The precise choice of ϵ_r will be determined later. Using the previous upper bound on Δ^*_{1,N_r}, the range of ϵ_r and with r large enough, we can write

$$\mathbf{P}_0(\Delta^*_{1,N_r} > 4\epsilon_r a_r)$$

$$\le \mathbf{P}_0\left(R_{N_r} > \epsilon_r a_r\right) + \mathbf{P}_0\left(\max_{|i| \le \sqrt{N_r}\log N_r} H_{N_r}(i) > \epsilon_r a_r\right)$$

$$\le \sum_{|i| \le \sqrt{N_r}\log N_r} \mathbf{P}_0\left(\left|\sum_{j=1}^{N_r} \xi_{j(i)}\right| > N_r \epsilon_r a_r\right) + \mathbf{P}_0\left(\left|\sum_{j=1}^{N_r} \xi'_{j(i)}\right| > N_r \epsilon_r a_r\right)$$

$$= K_1 + K_2, \quad \text{(say)}, \tag{6}$$

where,

$$\xi_{j(i)} = \mathbf{1}\left(\xi_p + \frac{i}{N_r} < X_j \le \xi_p + \frac{i}{N_r} + ta_r\right) - p_i,$$

and

$$\xi'_{j(i)} = \mathbf{1}\left(\xi_p + \frac{i-1}{N_r} < X_j \le \xi_p + \frac{i}{N_r}\right) - p'_i,$$

with $p_i = \mathbf{E}_0(\xi_{1(i)})$ and $p'_i = \mathbf{E}_0(\xi'_{1(i)})$. Thus, for a fixed i, $\{\xi_{j(i)} : j = 1, \ldots, N_r\}$ and $\{\xi'_{j(i)} : j = 1, \ldots, N_r\}$ are iid random variables, with the following distribution,

$$\xi_{j(i)} = \begin{cases} 1 - p_i & : \text{ w.p } p_i \\ p_i & : \text{ w.p } 1 - p_i \end{cases}$$

and exactly same for $\xi'_{j(i)}$ with p'_i. In order to find bounds on K_1 and K_2 in (6), we will use the inequality derived in Singh (1981) (in Lemma 1 above) on the variables $\xi_{j(i)}$ and $\xi'_{j(i)}$ respectively. Using the fact that $\epsilon_r \downarrow 0$, it can be shown after a careful choice of constants involved in Lemma 1, that

$$K_1 \le 4\sqrt{N_r} \log N_r \exp\left(-c_2|t|^{-1}N_r\epsilon_r^2 a_r\right). \tag{7}$$

And similarly, using Lemma 1 on the $\xi'_{j(i)}$ we can obtain the following bound on K_2,

$$K_2 \le 4\sqrt{N_r} \log N_r \exp(-c_3 N_r\epsilon_r a_r). \tag{8}$$

Here c_2 are c_3 are positive constants not depending on r. Now we choose the sequence ϵ_r as

$$\epsilon_r^3 = \frac{\log N_r}{N_r a_r} = \frac{\log N_r}{\sqrt{N_r(1 - f_r)}}\sqrt{f_r}.$$

Using the condition (1) we can say that $\epsilon_r \to 0$. Also note that $\epsilon_r^2 N_r a_r = \epsilon_r^{-1} \log N_r$. Using the relations (6)–(8) we can write,

$$\sum_{r=1}^{\infty} \mathbf{P}_0\left(\Delta^*_{1,N_r} > 4\epsilon_r a_r\right) \le \sum_{r=1}^{\infty} 4\sqrt{N_r} \log N_r \left[N_r^{(-c_2|t|^{-1}\epsilon_r^{-1})} + N_r^{(-c_3\epsilon_r^{-2})}\right] < \infty,$$

because $\epsilon_r^{-1} \to \infty$ as $r \to \infty$. Using Borel-Cantelli lemma we have,

$$\mathbf{P}_0\left(\frac{\Delta^*_{1,N_r}}{a_r} > 4\epsilon_r \ i.o.\right) = 0.$$

This implies, $\Delta^*_{1,N_r} = o(a_r) \ a.s. \ (\mathbf{P}_0)$. We can also conclude from the previous steps, that $R_{N_r} = o(a_r) \ a.s. \ (\mathbf{P}_0)$ and

$$\max_{|i| \le \sqrt{N_r} \log N_r} |\mathrm{H}_{N_r}(i)| = o(a_r) \ a.s. \ (\mathbf{P}_0).$$

Using similar methods it follows that $\Delta_{2,N_r}^* = o(a_r)$ a.s. ($\mathbf{P}_0$). This completes the proof of the lemma. $\qquad\square$

Lemma 3 *For each fixed $t \in \mathbb{R}$, $t \neq 0$, define the quantities*

$$p_{r,t} = 1 - q_{r,t} = F_{N_r}\left(\xi_{p,N_r} + ta_r\right) \quad and \quad c_{r,t} = \frac{\sqrt{n_r}\left(p_{r,t} - p\right)}{\sqrt{p_{r,t}(1 - p_{r,t})(1 - f_r)}}. \quad (9)$$

Assume that (1) *holds. Then, for all $1 \leq |t| \leq \log N_r$, there exists an $r_0 \in \mathbb{N}$ such that for all $r \geq r_0$,*

$$|c_{r,t}| \geq a|t|,$$

with $a = \frac{f_0(\xi_p)}{2}$ and $c_{r,t}$ is as defined above in (9).

Proof Using the triangle inequality on (9) we can write (please refer to (12) for definitions of $I_{1r,t}$, $I_{2r,t}$ and $I_{3r,t}$),

$$|c_{r,t}| \geq \frac{|t|}{\sqrt{p_{r,t}q_{r,t}}}\left(|I_{1r,t}| - |I_{2r,t}| - |I_{3r,t}|\right).$$

Note that for all $|t| \leq \log N_r$, $|ta_r| \to 0$. Thus we can use (13) to say that for large enough r and all $|t| \leq \log N_r$,

$$|I_{1r,t}| \geq \frac{f_0(\xi_p)}{2} \quad (> 0).$$

Using Lemma (4) we can say that $\sup\{|I_{2r,t}| : 1 \leq |t| \leq \log N_r\} \to 0$ a.s. ($\mathbf{P}_0$). And, since $(N_r a_r)^{-1} \downarrow 0$, we can say that $|I_{3r,t}| \to 0$ for all $|t| \in [1, \log N_r]$. Hence for a large r we can make $(|I_{2r,t}| + |I_{3r,t}|) < \frac{f_0(\xi_p)}{4}$ in the range $1 \leq |t| \leq \log N_r$. We also use the fact that for any t, $p_{r,t}(1 - p_{r,t}) \xrightarrow{a.s.} p(1 - p) \leq \frac{1}{4}$, so for r large ($>r_2$),

$$|c_{r,t}| \geq \frac{f_0(\xi_p)}{2}|t|,$$

which completes the proof of the lemma. $\qquad\square$

Lemma 4

$$\sup_{|t|\leq \log N_r} |F_{N_r}\left(\xi_{p,N_r} + ta_r\right) - F_{N_r}\left(\xi_{p,N_r}\right) - F_0\left(\xi_{p,N_r} + ta_r\right) + F_0\left(\xi_{p,N_r}\right)| = o\left(a_r\right),$$

$$(10)$$

almost surely ($\mathbf{P}_0$).

Proof The proof can be carried out in similar lines as in Lemma 2 and hence is omitted. $\qquad\square$

5.2 Proofs of the main results

Proof (*Theorem* 1) In this proof we consider t as a fixed real number. Let $Z_{j,t} = \mathbf{1}(Y_j \leq \xi_{p,N_r} + ta_r)$ for $j = 1, \ldots, n_r$. Then $\mathbf{E}_{\cdot|\mathcal{X}_{N_r}}(Z_{j,t}) = p_{r,t}$ (cf. 9). Then, the cdf of U_r can be written as

$$G_{n_r}(t) = 1 - \mathbf{P}_{\cdot|\mathcal{X}_{N_r}}\left(\frac{\sum_{j=1}^{n_r}(Z_{j,t} - n_r p_{r,t})}{\sqrt{n_r p_{r,t}(1 - p_{r,t})(1 - f_r)}} < -c_{r,t}\right), \tag{11}$$

Note that $\sum_{j=1}^{n_r} Z_{j,t} \sim Hyp(n_r; N_r p_{r,t}, N_r)$, where the pmf of a random variable $X \sim Hyp(n; M, N)$ is given as

$$P(X = x) \equiv P(x; n, M, N) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{if } x = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

Using (9) we can write,

$$c_{r,t} = \frac{t}{\sqrt{p_{r,t}(1 - p_{r,t})}}\left[I_{1r,t} + I_{2r,t} + I_{3r,t}\right], \tag{12}$$

where,

$$I_{1r,t} = \frac{F_0(\xi_{p,N_r} + ta_r) - F_0(\xi_{p,N_r})}{ta_r},$$

$$I_{2r,t} = \left\{\frac{F_{N_r}(\xi_{p,N_r} + ta_r) - F_{N_r}(\xi_{p,N_r}) - F_0(\xi_{p,N_r} + ta_r) + F_0(\xi_{p,N_r})}{ta_r}\right\},$$

and

$$I_{3r,t} = \frac{1}{t} O\left(\frac{1}{N_r a_r}\right).$$

We are interested in finding the limiting value of $c_{r,t}$. Using the Mean-value theorem, we can write

$$I_{1r,t} = \frac{F_0(\xi_{p,N_r} + ta_r) - F_0(\xi_{p,N_r})}{ta_r} = f_0(\xi_{p,N_r} + \theta_r ta_r), \quad (0 < \theta_r < 1).$$

Since t is fixed, we can use $\xi_{p,N_r} \to \xi_p$ a.s. $(\mathbf{P}_0)$, $a_r \downarrow 0$ and continuity of f_0 in a neighbourhood of ξ_p to say

$$I_{1r,t} = f_0(\xi_{p,N_r} + \theta_r ta_r) \longrightarrow f_0(\xi_p) \quad a.s. \ (\mathbf{P}_0). \tag{13}$$

Using the result in Lemma 2 we can say, $I_{2r,t} \to 0$ $a.s.$ ($\mathbf{P}_0$). And by (1), $(N_r a_r)^{-1} \to 0$, which in turn implies $I_{3r,t} \to 0$. Combining these we find that

$$\frac{p_{r,t} - p}{t a_r} = \left[I_{1r,t} + I_{2r,t} + I_{3r,t} \right] \longrightarrow f_0(\xi_p) \quad a.s. \ (\mathbf{P}_0).$$

As t is fixed we can use (12) and the above limit to conclude that $p_{r,t} \xrightarrow{a.s.} p$. This gives

$$c_{r,t} \longrightarrow \frac{t}{\sqrt{pq}} f_0(\xi_p) \quad a.s. \ (\mathbf{P}_0). \tag{14}$$

Using (11), write

$$\left| G_{n_r}(t) - \Phi(t/\rho) \right|$$

$$\leq \left| \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(\frac{\sum_{j=1}^{n_r} \left(Z_{j,t} - n_r p_{r,t} \right)}{\sqrt{n_r p_{r,t}(1 - p_{r,t})(1 - f_r)}} < -c_{r,t} \right) - \Phi\left(-c_{r,t} \right) \right|$$

$$+ \left| \Phi\left(c_{r,t} \right) - \Phi(t/\rho) \right|$$

$$= A_{1,r} + A_{2,r}, \quad \text{(say)}.$$

Since (1) holds and $p_{r,t} \to p$ almost surely (and $p \in (0, 1)$), we can say $\sigma_{r,t}^2 = \mathbf{V}_{\cdot | \mathcal{X}_{N_r}} (\sum_{j=1}^{n_r} Z_{j,t}) = n_r p_{r,t}(1 - p_{r,t})(1 - f_r) \to \infty$. This ensures that we can use Theorem 2.2 of Lahiri and Chatterjee (2007), which gives the following Berry-Esseen bound for Hypergeometric probabilities to bound the first term $A_{1,r}$ (solely under the condition that $\sigma_{r,t}^2 \to \infty$). Thus we can write

$$\sup_{u \in \mathbb{R}} \left| \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(\frac{\sum_{j=1}^{n_r} \left(Z_{j,t} - n_r p_{r,t} \right)}{\sqrt{n_r p_{r,t}(1 - p_{r,t})(1 - f_r)}} \leq u \right) - \Phi(u) \right| \leq \frac{C_1}{\sqrt{n_r p_{r,t} q_{r,t}(1 - f_r)}},$$

where C_1 is some positive constant that does not depend on r. This ensures that $A_{1,r} \to 0$ $a.s.$ ($\mathbf{P}_0$) as $r \to \infty$ for every $t \in \mathbb{R}$. We use (14) and the continuity of $\Phi(\cdot)$ to conclude that $A_{2,r} \to 0$ $a.s.$ ($\mathbf{P}_0$). Now restricting t over a countable dense subset of $\mathbb{R}$, it follows that $\mathcal{L}\left(U_r | \mathcal{X}_{N_r} \right) \xrightarrow{d} N(0, \rho^2)$ as $r \to \infty$, $a.s.$ ($\mathbf{P}_0$). In view of Polya's theorem this completes the proof of both assertions in Theorem 1.   □

Proof (*Theorem* 2) In order to prove the theorem, it will be enough to show

$$\sup_{r \in \mathbb{N}} \int_1^\infty t^{1+\delta} \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(a_r^{-1} |\widehat{\xi}_{p,n_r} - \xi_{p,N_r}| > t \right) dt < \infty. \tag{15}$$

Using the moment condition in Theorem 2 and Lemma 3 from Ghosh et al. (1984), applied to the random variables $\{X_1, \ldots, X_{N_r}\}$, we can write, $\max\{|X_j| : 1 \leq j \leq N_r\} = o(N_r^{1/\alpha})$ with probability 1. Since,

$$\max\{|Y_j| : 1 \leq j \leq n_r\} \ \leq \ \max\{|X_j| : 1 \leq j \leq N_r\},$$

with probability 1, and we can say that

$$\mathbf{P}_0 \left(\int_{N_r^{1/\alpha}}^{\infty} t^{1+\delta} \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(a_r^{-1} |\widehat{\xi}_{p,n_r} - \xi_{p,N_r}| > t \right) dt \neq 0 \; i.o. \right) = 0. \qquad (16)$$

Writing $[1, N_r^{1/\alpha}] = [1, \sqrt{\log N_r}) \cup [\sqrt{\log N_r}, N_r^{1/\alpha}]$, we have

$$\int_1^{N_r^{1/\alpha}} t^{1+\delta} \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(a_r^{-1} |\widehat{\xi}_{p,n_r} - \xi_{p,N_r}| > t \right) dt$$

$$\leq \int_1^{\sqrt{\log N_r}} + \frac{N_r^{\frac{2+\delta}{\alpha}}}{2+\delta} \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(a_r^{-1} |\widehat{\xi}_{p,n_r} - \xi_{p,N_r}| > \sqrt{\log N_r} \right)$$

$$\equiv B_{1,r} + B_{2,r} \quad \text{(say)}.$$

Using (9) and (11), we have

$$\mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(|\widehat{\xi}_{p,n_r} - \xi_{p,N_r}| > t a_r \right)$$
$$\leq \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(|Z_{n_r,t}| > c_{r,t} \right) + \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(|Z_{n_r,-t}| > c_{r,-t} \right),$$

where

$$Z_{n_r,t} = \frac{\sum_{j=1}^{n_r} \left(Z_{j,t} - n_r p_{r,t} \right)}{\sqrt{n_r p_{r,t}(1 - p_{r,t})(1 - f_r)}}.$$

We will use Corollary 2.4 of Lahiri et al. (2006) which gives an exponential upper bound on the tail probabilities of a Hypergeometric random variable to bound the tail probabilities in the above expression. *The bound states that*

$$\mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(|Z_{n_r,t}| \geq u \right) \leq \frac{C_4}{\left(p_{r,t} \wedge q_{r,t} \right)^3} \exp \left(-C_5 u^2 \left[p_{r,t} \wedge q_{r,t} \right]^2 \right), \quad u \in \mathbb{R}, \quad (17)$$

where, C_4, C_5 are universal constants not depending on r and t is a fixed real number and $x \wedge y = \min\{x, y\}$. Define $p_{r,j} = 1 - q_{r,j}$, $j = 1, 2$ using $t = \pm \sqrt{\log N_r}$ in (9),

$$p_{r,1} = F_{N_r} \left(\xi_{p,N_r} + \sqrt{\log N_r} a_r \right) \quad \text{and} \quad p_{r,2} = F_{N_r} \left(\xi_{p,N_r} - \sqrt{\log N_r} a_r \right).$$

Using (17) at $t = \pm \sqrt{\log N_r}$, we can have the following

$$B_{2,r} \leq \frac{C_4 N_r^{\frac{2+\delta}{\alpha}}}{2+\delta} \left[\frac{N_r^{-\left(C_5 [p_{r,1} \wedge q_{r,1}]^2 \right)}}{\left(p_{r,1} \wedge q_{r,1} \right)^3} + \frac{N_r^{-\left(C_5 [p_{r,2} \wedge q_{r,2}]^2 \right)}}{\left(p_{r,2} \wedge q_{r,2} \right)^3} \right]. \qquad (18)$$

Using the same arguments that we used in proving $p_{r,t} \xrightarrow{a.s.} p$ (see (14)), we can say that $\left(p_{r,j} \wedge q_{r,j} \right) \xrightarrow{a.s.} (p \wedge q)$ as $r \to \infty$ for each $j = 1, 2$. Since $(p \wedge q) > 0$,

there exists some random $R_1 \in \mathbb{N}$ (depending on the selected sample $\mathcal{X}_{N_r}$), such that $0 < (p \wedge q)/2 < (p_{r,j} \wedge q_{r,j})$ for all $r \geq R_1$ for each $j = 1, 2$. Hence each term on the right side of (18) can be bounded, for all $r \geq R_1$, by

$$\frac{N_r^{\frac{2+\delta}{\alpha}}}{\left(p_{r,j} \wedge q_{r,j}\right)^3} N_r^{-\left(C_5[p_{r,j} \wedge q_{r,j}]^2\right)} \leq \frac{8 N_r^{\frac{2+\delta}{\alpha}}}{(p \wedge q)^3} N_r^{-C_5(k_0[p \wedge q]^2)/4} \quad \forall j = 1, 2,$$

where, $k_0 \in (0, \infty)$ is an appropriately chosen constant depending only on δ, α and p such that $B_{2,r} \downarrow 0$. For bounding the term in $B_{1,r}$ we use Lemma 3 in the range $t \in [1, \sqrt{\log N_r}]$ to write

$$\mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(|Z_{n_r,t}| > c_{r,t} \right) \leq \mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(|Z_{n_r,t}| > a|t| \right),$$

where $a = \frac{f_0(\xi_p)}{2}$. Using the exponential bound (17) along with this inequality, we can have the following upper bound

$$\mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(a_r^{-1} |\widehat{\xi}_{p,n_r} - \xi_{p,N_r}| > t \right) dt \leq K_1 \exp(-K_2 t^2) \quad \forall t \in [1, \sqrt{\log N_r}],$$

for some positive constants K_1 and K_2. Thus for any choice of $\delta > 0$, the integral in (15) is finite. This shows that $\{U_r^2 : r \geq 1\}$ is uniformly integrable as claimed in (2). $\square$

Proof (*Corollary* 1) Using the results of Theorems 1 and 2, the proof of the Corollary follows. $\square$

Proof (*Theorem* 3) The result follows directly from the proof of Theorem 4(a) below. $\square$

Proof (*Theorem* 4(a)) Define as earlier (cf. (9))

$$\tilde{p}_{r,t} = \widehat{F}_{n_r}(\widehat{\xi}_{p,n_r} + t b_r) \quad \text{and} \quad d_{r,t} = \frac{\sqrt{n_r}(\tilde{p}_{r,t} - p)}{\sqrt{\tilde{p}_{r,t}(1 - \tilde{p}_{r,t})}}.$$

Also define, $V'_{j,t} = \mathbf{1}\left(Y_j^* \leq \widehat{\xi}_{p,n_r} + t b_r\right)$ for $j = 1, \dots, n_r$, with $\mathbf{E}_*(V'_{j,t}) = \tilde{p}_{r,t}$. Then, the d.f. of U_r^{**} can be written as

$$G_{n_r}^*(t) = 1 - \mathbf{P}_* \left(\frac{\sum_{j=1}^{n_r}(V'_{j,t} - \tilde{p}_{r,t})}{\sqrt{n_r \tilde{p}_{r,t}(1 - \tilde{p}_{r,t})}} < -d_{r,t} \right).$$

The $\{Y_j^*\}$ are *iid* random variables and $\sum_{j=1}^{n_r} V'_{j,t} \sim bin(n_r, \tilde{p}_{r,t})$. We can use the Berry–Esseen theorem for *iid* random variables to write

$$\sup_{u \in \mathbb{R}} \left| \mathbf{P}_* \left(\frac{\sum_{j=1}^{n_r} V'_{j,t} - n_r \tilde{p}_{r,t}}{\sqrt{n_r \tilde{p}_{r,t}(1 - \tilde{p}_{r,t})}} \leq u \right) - \Phi(u) \right| \leq \frac{C_0}{\sqrt{n_r \tilde{p}_{r,t}(1 - \tilde{p}_{r,t})}}, \quad (19)$$

for some finite C_0. Now consider the limiting value of $d_{r,t}$. The numerator of $d_{r,t}$ can be written as,

$$\sqrt{n_r}\left(\tilde{p}_{r,t} - p\right) = J_{1r,t} + J_{2r,t} + O\left(n_r^{-1/2}\right), \tag{20}$$

where,

$$J_{1r,t} = \sqrt{n_r}\left(F_{N_r}(\widehat{\xi}_{p,n_r} + tb_r) - F_{N_r}(\widehat{\xi}_{p,n_r})\right)$$

and,

$$J_{2r,t} = \sqrt{n_r}\left(\widehat{F}_{n_r}(\widehat{\xi}_{p,n_r} + tb_r) - F_{N_r}(\widehat{\xi}_{p,n_r} + tb_r) + F_{N_r}(\widehat{\xi}_{p,n_r}) - \widehat{F}_{n_r}(\widehat{\xi}_{p,n_r})\right).$$

We can split $J_{1r,t}$ into two parts as follows:

$$\begin{aligned}
J_{1r,t} &= \sqrt{n_r}\left[F_0(\widehat{\xi}_{p,n_r} + tb_r) - F_0(\widehat{\xi}_{p,n_r})\right] \\
&\quad + \sqrt{n_r}\left[F_{N_r}(\widehat{\xi}_{p,n_r} + tb_r) - F_0(\widehat{\xi}_{p,n_r} + tb_r) + F_0(\widehat{\xi}_{p,n_r}) - F_{N_r}(\widehat{\xi}_{p,n_r})\right] \\
&= J'_{1r,t} + J''_{1r,t} \quad \text{(say)}.
\end{aligned}$$

For all $r \geq 1$, define the set

$$S_r = \left\{|\widehat{\xi}_{p,n_r} - \xi_p| < \frac{2\log n_r}{\sqrt{n_r}}\right\}. \tag{21}$$

Then,

$$\begin{aligned}
\mathbf{P}_0(S_r^c) &\leq \mathbf{E}_0\left(\mathbf{P}_{\cdot|\mathcal{X}_{N_r}}\left(\sqrt{n_r}|\widehat{\xi}_{p,n_r} - \xi_{p,N_r}| > \log n_r\right)\right) \\
&\quad + \mathbf{P}_0(\sqrt{n_r}|\xi_{p,N_r} - \xi_p| > \log n_r).
\end{aligned}$$

The first term in the above bound can be shown to converge to zero, by using Theorem 1 along with DCT. The second term converges to zero and that can be shown by the CLT for sample quantiles for sampling from a continuous population. This implies

$$\mathbf{P}_0(S_r^c) \to 0 \quad \text{as } r \to \infty. \tag{22}$$

Using the mean-value theorem we can write

$$J'_{1r,t} = \frac{\sqrt{1 - f_r}}{\sqrt{1 - f}} t f_0(\widehat{\xi}_{p,n_r} + \theta t b_r), \tag{23}$$

for some $\theta \in (0, 1)$. On the set S_r, we can write $\widehat{\xi}_{p,n_r} = \xi_p + \frac{y_r}{\sqrt{n_r}}$, where $|y_r| \leq \log n_r$. Hence on the set S_r, the right side of (23) can be written as $f_0(\xi_p + \frac{y_r}{\sqrt{n_r}} + \theta t b_r)$. As

t is fixed, $f_r \to f, b_r \to 0, \frac{\log n_r}{\sqrt{n_r}} \downarrow 0$ and f_0 is continuous in a neighbourhood of ξ_p, we can write for any $\epsilon > 0$ (and on the set S_r),

$$|J'_{1r,t} - t f_0(\xi_p)| < \epsilon,$$

for large enough r. Hence, as $r \to \infty$,

$$\mathbf{P}_0 \left(|J'_{1r,t} - t f_0(\xi_p)| > \epsilon \right) \leq \mathbf{P}_0 \left(\{ |J'_{1r,t} - t f_0(\xi_p)| > \epsilon \} \cap S_r \right) + \mathbf{P}_0(S_r^c) \to 0.$$

So, $J'_{1r,t} \xrightarrow{\mathbf{P}_0} t f_0(\xi_p)$ as $r \to \infty$. Using the same argument (on the sets S_r) as used in proving Lemma 2, we can say that $J''_{1r,t} \xrightarrow{\mathbf{P}_0} 0$ as $r \to \infty$. This implies that

$$J_{1r,t} \xrightarrow{\mathbf{P}_0} t f_0(\xi_p) \quad \text{as } r \to \infty. \tag{24}$$

Now we can write

$$J_{2r,t} = \frac{1}{\sqrt{n_r}} \sum_{j=1}^{n_r} \left(W_{j,t} - \mathbf{E}_{\cdot | \mathcal{X}_{N_r}} (W_{j,t}) \right), \tag{25}$$

where $W_{j,t} = \mathbf{1} \left(\widehat{\xi}_{p,n_r} < Y_j \leq \widehat{\xi}_{p,n_r} + t b_r \right), j = 1, \ldots, n_r$. Note that $W_{j,t} = 1$ or 0 with probabilities $\omega_{r,t} = (F_{N_r}(\widehat{\xi}_{p,n_r} + t b_r) - F_{N_r}(\widehat{\xi}_{p,n_r}))$ and $(1 - \omega_{r,t})$ respectively. And $\sum_{j=1}^{n_r} W_{j,t} \sim Hyp \ (n_r; \ N_r \omega_{r,t}, \ N_r)$. On the set S_r, we can write $W_{j,t} = \mathbf{1}(\xi_p + \frac{y_r}{\sqrt{n_r}} < Y_j \leq \xi_p + \frac{y_r}{\sqrt{n_r}} + t b_r), j = 1, \ldots, n_r$. Now using Chebyshev's inequality, for any $\epsilon > 0$, we have

$$\mathbf{P}_{\cdot | \mathcal{X}_{N_r}} \left(\{ |J_{2r,t}| \geq \epsilon \} \cap S_r \right) \leq \frac{\tilde{\omega}_{r,t}(1 - \tilde{\omega}_{r,t})(1 - f_r)}{\epsilon^2},$$

where $\tilde{\omega}_{r,t} = \left(F_{N_r} \left(\xi_p + \frac{y_r}{\sqrt{n_r}} + t b_r \right) - F_{N_r} \left(\xi_p + \frac{y_r}{\sqrt{n_r}} \right) \right)$ is the value of $\omega_{r,t}$ on the set S_r. Using earlier arguments as used in proving (24) we can say $\tilde{\omega}_{r,t} \to 0$ on the set S_r. Using (22) and DCT it follows that

$$J_{2r,t} \xrightarrow{\mathbf{P}_0} 0.$$

Using these results we can say $\tilde{p}_{r,t} \xrightarrow{\mathbf{P}_0} p \ (\in (0, 1))$. This implies

$$d_{r,t} \xrightarrow{\mathbf{P}_0} \frac{t f_0(\xi_p)}{\sqrt{p(1-p)}} \quad \text{as } r \to \infty. \tag{26}$$

Thus we have

$$
\left| G_{n_r}^*(t) - \Phi(t/\rho) \right| \leq \left| \mathbf{P}_* \left(\frac{\sum_{j=1}^{n_r} V_{j,t}' - n_r \tilde{p}_{r,t}}{\sqrt{n_r \tilde{p}_{r,t}(1 - \tilde{p}_{r,t})}} < -d_{r,t} \right) - \Phi(-d_{r,t}) \right|
$$
$$
+ |\Phi(d_{r,t}) - \Phi(t/\rho)|.
$$

Now, using (19), the continuity of $\Phi(.)$ and (26) we can say that for all $t \in \mathbb{R}$, $G_{n_r}^*(t) \xrightarrow{\mathbf{P_0}} \Phi(t/\rho)$ as $r \to \infty$. This completes the proof of Theorem 4(a).   □

Proof (*Theorem* 4b)   It will be enough to show that

$$
\sup_{r \geq 1} \int_1^\infty t^{1+\delta} \mathbf{P}_*(|U_r^{**}| \geq t) \mathrm{d}t, \tag{27}
$$

is bounded for some choice of $\delta > 0$. By our earlier argument (see 16) we can say that

$$
\int_{n_r^{1/\alpha}}^\infty t^{1+\delta} \mathbf{P}_*(|U_r^{**}| \geq t) \mathrm{d}t \to 0 \quad a.s. \ (\mathbf{P_0}) \quad \text{as } r \to \infty.
$$

We consider the term, $\mathbf{P}_*(U_r^{**} \geq t)$ and split the range of the integral in (27) into two parts: $[1, n_r^{1/\alpha}) = [1, \log n_r] \cup (\log n_r, n_r^{1/\alpha})$.

$$
\int_1^{n_r^{1/\alpha}} t^{1+\delta} \mathbf{P}_*(U_r^{**} \geq t) \, \mathrm{d}t = \int_1^{\log n_r} + \int_{\log n_r}^{n_r^{1/\alpha}} t^{1+\delta} \mathbf{P}_*(U_r^{**} \geq t) \, \mathrm{d}t. \tag{28}
$$

We can rewrite $\mathbf{P}_*(U_r^{**} \geq t) = \mathbf{P}_* \left(\bar{V}_{n_r,t}' - \tilde{p}_{r,t} < -(\tilde{p}_{r,t} - p) \right)$, where we define (for notational simplicity)

$$
\bar{V}_{n_r}'(i) \equiv \bar{V}_{n_r, \frac{i}{\sqrt{n_r}}}' = \frac{1}{n_r} \sum_{j=1}^{n_r} \mathbf{1} \left(Y_j^* \leq \widehat{\xi}_{p,n_r} + \frac{i}{\sqrt{n_r}} b_r \right)
$$

and,

$$
\tilde{p}_r(i) \equiv \tilde{p}_{r, \frac{i}{\sqrt{n_r}}} = \widehat{F}_{n_r} \left(\widehat{\xi}_{p,n_r} + \frac{i}{\sqrt{n_r}} b_r \right).
$$

Thus,

$$
\int_1^{\log n_r} t^{1+\delta} \mathbf{P}_*(U_r^{**} \geq t) \, \mathrm{d}t
$$
$$
\leq \sum_{i=1}^{\lfloor \sqrt{n_r} \log n_r \rfloor} \left(\frac{i+1}{\sqrt{n_r}} \right)^{1+\delta} \mathbf{P}_* \left(\bar{V}_{n_r}'(i) - \tilde{p}_r(i) < -(\tilde{p}_r(i) - p) \right). \tag{29}
$$

From (20) we can write,

$$(\tilde{p}_r(i) - p) = \frac{1}{\sqrt{n_r}} \left[J'_{1r}(i) + J''_{1r}(i) + J_{2r}(i) \right] + O\left(n_r^{-1}\right),$$

where, $J'_{1r}(i)$, $J''_{1r}(i)$, $J_{2r}(i)$ are similarly defined as $J'_{1r,t}$, $J''_{1r,t}$ and $J_{2r,t}$ respectively, with $t = \frac{i}{\sqrt{n_r}}$. Also define $W_j(i)$ in a similar way as $W_{j,t}$ (cf. 25). On the set S_r, we can write

$$W_j(i) = \mathbf{1}\left(\xi_p + \frac{y_r}{\sqrt{n_r}} < Y_j \le \xi_p + \frac{y_r}{\sqrt{n_r}} + tb_r\right),$$

with $|y_r| \le \log n_r$. Now consider any $\epsilon > 0$, we can write using Hoeffding's inequality (on the $W_j(i)$'s) for sampling without replacement (see Hoeffding 1963, Sect. 6),

$$\mathbf{P}_{\cdot|\mathcal{X}_{N_r}}\left(\left\{\max_{1 \le i \le \lfloor \sqrt{n_r}\log n_r \rfloor} \frac{|J_{2r}(i)|}{\sqrt{n_r}} > \epsilon\right\} \cap S_r\right)$$

$$\le \sum_{i=1}^{\lfloor \sqrt{n_r}\log n_r \rfloor} \mathbf{P}_{\cdot|\mathcal{X}_{N_r}}\left(\left\{\left|\sum_{j=1}^{n_r}(W_j(i) - \mathbf{E}_{\cdot|\mathcal{X}_{N_r}}(W_j(i)))\right| > \epsilon n_r\right\} \cap S_r\right)$$

$$\le 2\exp(-2n_r\epsilon^2)\sqrt{n_r}\log n_r \to 0 \quad \text{as } r \to \infty.$$

Using (22), we can say

$$\max\left\{\frac{J_{2r}(i)}{\sqrt{n_r}} : 1 \le i \le \lfloor \sqrt{n_r}\log n_r \rfloor\right\} \xrightarrow{\mathbf{P}_0} 0. \tag{30}$$

Using Hoeffding's inequality (for sums of independent random variables) on the set S_r and using (22) we can say

$$\max\left\{\frac{J''_{1r}(i)}{\sqrt{n_r}} : 1 \le i \le \lfloor \sqrt{n_r}\log n_r \rfloor\right\} \xrightarrow{\mathbf{P}_0} 0. \tag{31}$$

Using (30) and (31) we can say

$$-(\tilde{p}_r(i) - p) \le -\frac{J'_{1r}(i)}{2\sqrt{n_r}},$$

uniformly for all $1 \le i \le \lfloor \sqrt{n_r}\log n_r \rfloor$. Now,

$$\frac{J'_{1r}(i)}{\sqrt{n_r}} = f_0\left(\widehat{\xi}_{p,n_r} + \theta_i \frac{ib_r}{\sqrt{n_r}}\right)\frac{ib_r}{\sqrt{n_r}},$$

for some $\theta_i \in (0, 1)$. Since $\max\{|\frac{ib_r}{\sqrt{n_r}}| : 1 \le i \le \lfloor \sqrt{n_r}\log n_r \rfloor\} \to 0$, hence using the same argument (on the set S_r) we can say that $f_0(\widehat{\xi}_{p,n_r} + \theta_i \frac{ib_r}{\sqrt{n_r}}) \to f_0(\xi_p)$ in probability $\mathbf{P}_0$. Thus on the set S_r, using Hoeffding's inequality we can write

$$\mathbf{P}_* \left(\bar{V}'_{n_r}(i) - \tilde{p}_r(i) < -(\tilde{p}_r(i) - p) \right) \le \mathbf{P}_* \left(\bar{V}'_{n_r}(i) - \tilde{p}_r(i) < -\frac{J'_{1r}(i)}{2\sqrt{n_r}} \right)$$

$$\le \mathbf{P}_* \left(|\bar{V}'_{n_r}(i) - \tilde{p}_r(i)| > \frac{f_0(\xi_p)}{4} \frac{i b_r}{\sqrt{n_r}} \right)$$

$$\le 2 \exp\left(-\frac{f_0^2(\xi_p)}{8} i^2 b_r^2 \right)$$

$$= 2 \exp(-K_3 i^2 b_r^2),$$

for some constant K_3 (independent of i). Using this bound in (29), we have

$$\int_1^{\log n_r} t^{1+\delta} \mathbf{P}_*(U_r^{**} \ge t) \, dt$$

$$\le \sum_{i=1}^{\lfloor \sqrt{n_r} \log n_r \rfloor} \left(\frac{i+1}{\sqrt{n_r}} \right)^{1+\delta} 2 \exp\left(-K_3(i b_r)^2 \right) = O(1), \tag{32}$$

which follows from the convergence of the series $\sum_{m=1}^{\infty} m^{1+\delta} \exp(-am^2)$, (for any $a, \delta > 0$). Using $i = \sqrt{n_r} \log n_r$, we can obtain the following bound for $t = \log n_r$,

$$\mathbf{P}_* \left(\bar{V}'_{n_r, \log n_r} - \tilde{p}_{r, \log n_r} < -(\tilde{p}_{r, \log n_r} - p) \right) \le 2 \exp\left(-2K_4 (\log n_r)^2 \right),$$

on the set S_r, where K_4 is some finite constant. Then on the set S_r we can write,

$$\int_{\log n_r}^{n_r^{1/\alpha}} t^{1+\delta} \mathbf{P}_*(U_r^{**} \ge t) \, dt \le K_5 n_r^{\frac{2+\delta}{\alpha}} \exp\left(-2K_4 (\log n_r)^2 \right) = o(1). \tag{33}$$

Using (32) and (33) we can say that the integral in (28) is finite. We can similarly deal with the integral involving $\mathbf{P}_*(U_r^{**} \le -t)$. Combining both, we can show that $\{|U_r^{**}|^2 : r \ge 1\}$ is uniformly integrable. Again using (22) and combining this with the asymptotic normality we have $\mathbf{V}_*(U_r^{**}) \to \rho^2$ in probability $\mathbf{P}_0$. □

References

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics, 37*, 577–580.

Bickel, P. J., Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics, 12*(2), 470–482.

Booth, J. G., Butler, R. W., Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association, 89*(428), 1282–1289.

Chambers, R. L., Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika, 73*, 597–604.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.

Francisco, C. A., Fuller, W. A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics, 19*(1), 454–469.

Ghosh, M., Parr, W. C., Singh, K., Babu, G. J. (1984). A note on bootstrapping the sample median. *The Annals of Statistics, 12*(3), 1130–1135.

Gross, S. T. (1980). Median estimation in sample surveys. In *ASA Proceedings of the Section on Survey Research Methods* (pp. 181–184).

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association, 58*, 13–30.

Isaki, C. T., Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association, 77*(377), 89–96.

Kovar, J. G., Rao, J. N. K., Wu, C.-F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics, 16*(suppl.), 25–45.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science, 18*(2), 199–210.

Lahiri, S. N., Chatterjee, A. (2007). A Berry-Esseen theorem for hypergeometric probabilities under minimal conditions. *Proceedings of the American Mathematical Society, 135*(5), 1535–1545 (electronic).

Lahiri, S. N., Chatterjee, A., Maiti, T. (2006). A sub-gaussian berry-esseen theorem for the hypergeometric distribution. preprint available at http://arxiv.org/abs/math/0602276.

Mak, T. K., Kuk, A. (1993). A new method for estimating finite-population quantiles using auxiliary information. *The Canadian Journal of Statistics, 21*, 29–38.

McCarthy, P. J. (1965). Stratified sampling and distribution-free confidence intervals for a median. *Journal of the American Statistical Association, 60*, 772–783.

Nelson, D., Meeden, G. (2006). Noninformative nonparametric quantile estimation for simple random samples. *Journal of Statistical Planning and Inference, 136*(1), 53–67.

Rao, J. N. K., Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association, 83*, 231–241.

Rao, J. N. K., Kovar, J. G., Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika, 77*, 365–375.

Sedransk, J., Meyer, J. (1978). Confidence intervals for the quantiles of a finite population: Simple random and stratified simple random sampling. *Journal of the Royal Statistical Society, Series B, Methodological, 40*, 239–252.

Shao, J. (1994). *L*-statistics in complex survey problems. *The Annals of Statistics, 22*(2), 946–967.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science, 18*(2), 191–198.

Shao, J., Chen, Y. (1998). Bootstrapping sample quantiles based on complex survey data under hot deck imputation. *Statistica Sinica, 8*, 1071–1086.

Shao, J., Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics, 17*(3), 1176–1197.

Shao, J., Wu, C.-F. J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *The Annals of Statistics, 20*(3), 1571–1593.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics, 9*(6), 1187–1195.

Sitter, R. R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics, 20*, 135–154.

Smith, P., Sedransk, J. (1983). Lower bounds for confidence coefficients for confidence intervals for finite population quantiles. *Communications in Statistics, Part A—Theory and Methods, 12*, 1329–1344.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association, 47*, 635–646.